Adaptive Optimization Methods for Machine Learning and Signal Processing

Volkan Cevher
volkan.cevher@epfl.ch

Ali Kavis
ali.kavis@epfl.ch

Kfir Y. Levy
kfirylevy@technion.ac.il

Ahmet Alacaoglu
ahmet.alacaoglu@epfl.ch

Part IV/IV: Adaptivity in min-max optimization

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

Setup: min-max optimization

$$\boxed{\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \Phi(x,y)}$$

- $\Phi(\cdot, y)$ is convex for all y
- $\Phi(x, \cdot)$ is concave for all x
- $\mathcal{X}, \mathcal{Y}$ are closed, convex sets
- Solution set is nonempty: $\exists x^\star, y^\star$:

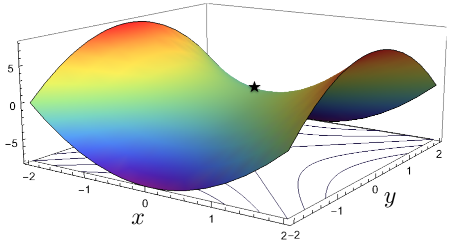
$$\Phi(x^\star, y^\star) = \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \Phi(x,y)$$

Setup: min-max optimization

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \Phi(x,y)$$

- $\Phi(\cdot,y)$ is convex for all y
- $\Phi(x,\cdot)$ is concave for all x
- $\mathcal{X},\mathcal{Y}$ are closed, convex sets
- Solution set is nonempty: $\exists x^\star, y^\star$:

$$\Phi(x^\star,y^\star) = \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \Phi(x,y)$$



- The solution is the saddle point

Outline

→ Classic methods

Adaptivity to	Lipschitz constant	noise	strong convexity	sparsity of data
[Bach and Levy, 2019]	✓	✓	×	×
[Chambolle et al., 2018]	×	×	✓	×
[Alacaoglu et al., 2020]	×	×	✓	✓

Gradient Descent-ascent

$$\boxed{\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \Phi(x, y)}$$

Algorithm Simultaneous GDA (Forward-Backward) [Sibony, 1970]

for $t = 0$ to $T-1$ **do**
  $x_{t+1} = P_{\mathcal{X}}(x_t - \eta \nabla_x \Phi(x_t, y_t))$
  $y_{t+1} = P_{\mathcal{Y}}(y_t + \eta \nabla_y \Phi(x_t, y_t))$
end for

Algorithm Alternating GDA (Arrow-Hurwicz) [Arrow et al., 1958]

for $t = 0$ to $T-1$ **do**
  $x_{t+1} = P_{\mathcal{X}}(x_t - \eta \nabla_x \Phi(x_t, y_t))$
  $y_{t+1} = P_{\mathcal{Y}}(y_t + \eta \nabla_y \Phi(\boxed{x_{t+1}}, y_t))$
end for

Gradient Descent-ascent

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \Phi(x,y)$$

Algorithm Simultaneous GDA (Forward-Backward) [Sibony, 1970]

for $t = 0$ to $T-1$ **do**
$x_{t+1} = P_{\mathcal{X}}(x_t - \eta\nabla_x\Phi(x_t, y_t))$
$y_{t+1} = P_{\mathcal{Y}}(y_t + \eta\nabla_y\Phi(x_t, y_t))$
end for

Algorithm Alternating GDA (Arrow-Hurwicz) [Arrow et al., 1958]

for $t = 0$ to $T-1$ **do**
$x_{t+1} = P_{\mathcal{X}}(x_t - \eta\nabla_x\Phi(x_t, y_t))$
$y_{t+1} = P_{\mathcal{Y}}(y_t + \eta\nabla_y\Phi(\boxed{x_{t+1}}, y_t))$
end for

- Behavior of GDA on the toy problem:

$$\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} xy$$

Theorem. [Gidel et al., 2018] With any $\eta > 0$, iterates of simGDA diverge:

$$x_{t+1}^2 + y_{t+1}^2 = (1+\eta^2)(x_t^2 + y_t^2)$$

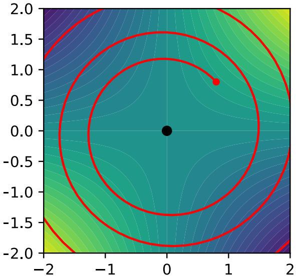
Theorem. [Gidel et al., 2018] With any $\eta > 0$, iterates of altGDA do not converge:

$$x_{t+1}^2 + y_{t+1}^2 = \Theta(x_0^2 + y_0^2)$$

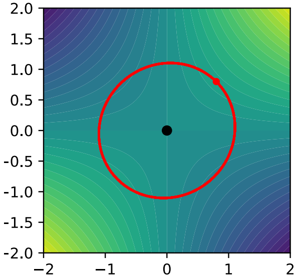
Toy problem: in practice

$$\boxed{\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} xy}$$

- Simultaneous GDA



- Alternating GDA



Extragradient

$$\boxed{\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\Phi(x,y)}$$

◦ Notation for convenience: $z=(x,y)$, $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$, $F(z)=(\nabla_x\Phi(x,y),-\nabla_y\Phi(x,y))$.

Algorithm Simultaneous GDA (Forward-backward)

for $t=0$ to $T-1$ **do**
  $z_{t+1}=P_{\mathcal{Z}}(z_t-\eta F(z_t))$
end for

◦ Note the equivalence to

$$x_{t+1}=P_{\mathcal{X}}(x_t-\eta\nabla_x\Phi(x_t,y_t))$$
$$y_{t+1}=P_{\mathcal{Y}}(y_t+\eta\nabla_y\Phi(x_t,y_t))$$

Extragradient

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \Phi(x,y)$$

- Notation for convenience: $z=(x,y)$, $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$, $F(z)=(\nabla_x\Phi(x,y), -\nabla_y\Phi(x,y))$.

Algorithm Simultaneous GDA (Forward-backward)

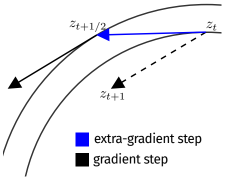
for $t=0$ to $T-1$ **do**
  $z_{t+1}=P_{\mathcal{Z}}(z_t-\eta F(z_t))$
end for

- Note the equivalence to

$$x_{t+1}=P_{\mathcal{X}}(x_t-\eta\nabla_x\Phi(x_t,y_t))$$
$$y_{t+1}=P_{\mathcal{Y}}(y_t+\eta\nabla_y\Phi(x_t,y_t))$$

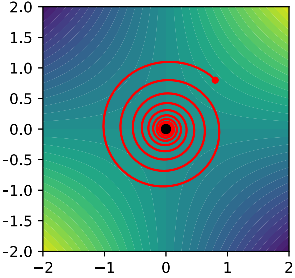
Algorithm Extragradient [Korpelevich, 1976]

for $t=0$ to $T-1$ **do**
  $z_{t+1/2}=P_{\mathcal{Z}}(z_t-\eta F(z_t))$
  $z_{t+1}=P_{\mathcal{Z}}(z_t-\eta F(z_{t+1/2}))$
end for



Toy problem: in practice

$$\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} xy$$



Convergence of Extragradient

$$\boxed{\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \Phi(x,y)}$$

- Notation for convenience: $z=(x,y)$, $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$, $F(z)=(\nabla_x\Phi(x,y),-\nabla_y\Phi(x,y))$.

Algorithm (Stochastic) Extragradient[1]

for $t=0$ to $T-1$ **do**
  $z_{t+1/2}=P_{\mathcal{Z}}(z_t-\eta_t\tilde{F}(z_t))$
  $z_{t+1}=P_{\mathcal{Z}}(z_t-\eta_t\tilde{F}(z_{t+1/2}))$
end for

Assumptions.

- $\Phi(x,y)$ is convex-concave
- A solution $z^\star$ exists
- Lipschitzness: $\|F(u)-F(v)\|\le L\|u-v\|$
- Unbiasedness: $\mathbb{E}[\tilde{F}(z)]=F(z)$
- Variance bound: $\mathbb{E}\|\tilde{F}(z)-F(z)\|^2\le R^2$

[1]Same ideas also apply when we use Bregman distances in the update rule. This version is known as Mirror-Prox.
[2]Merit function for the rate is primal-dual gap: $\mathrm{err}(\bar{z})=\max_{x,y}\Phi(\bar{x},y)-\Phi(x,\bar{y})$.

Convergence of Extragradient

$$\boxed{\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \Phi(x,y)}$$

- Notation for convenience: $z = (x,y)$, $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$, $F(z) = (\nabla_x\Phi(x,y), -\nabla_y\Phi(x,y))$.

Algorithm (Stochastic) Extragradient[1]

for $t = 0$ to $T-1$ **do**
  $z_{t+1/2} = P_{\mathcal{Z}}(z_t - \eta_t \tilde{F}(z_t))$
  $z_{t+1} = P_{\mathcal{Z}}(z_t - \eta_t \tilde{F}(z_{t+1/2}))$
end for

Assumptions.

- $\Phi(x,y)$ is convex-concave
- A solution $z^\star$ exists
- Lipschitzness: $\|F(u) - F(v)\| \le L\|u-v\|$
- Unbiasedness: $\mathbb{E}[\tilde{F}(z)] = F(z)$
- Variance bound: $\mathbb{E}\|\tilde{F}(z) - F(z)\|^2 \le R^2$

Theorem (Deterministic). [Korpelevich, 1976, Nemirovski, 2004]

EG w/ $\tilde{F} = F$, $\boxed{\eta_t = \eta < 1/L}$:

- Convergence: $z_{t+1} \to z^\star$
- Rate[2]: $\mathrm{err}\left(\frac{1}{T}\sum_{t=1}^{T} z_{t+1/2}\right) \le \mathcal{O}\left(\frac{1}{T}\right)$

[1] Same ideas also apply when we use Bregman distances in the update rule. This version is known as Mirror-Prox.
[2] Merit function for the rate is primal-dual gap: $\mathrm{err}(\bar{z}) = \max_{x,y} \Phi(\bar{x}, y) - \Phi(x, \bar{y})$.

Convergence of Extragradient

$$\boxed{\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}} \Phi(x,y)}$$

- Notation for convenience: $z = (x,y)$, $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$, $F(z) = (\nabla_x\Phi(x,y), -\nabla_y\Phi(x,y))$.

Algorithm (Stochastic) Extragradient[1]

for $t = 0$ to $T-1$ **do**
  $z_{t+1/2} = P_{\mathcal{Z}}(z_t - \eta_t\tilde{F}(z_t))$
  $z_{t+1} = P_{\mathcal{Z}}(z_t - \eta_t\tilde{F}(z_{t+1/2}))$
end for

Assumptions.

- $\Phi(x,y)$ is convex-concave
- A solution $z^\star$ exists
- Lipschitzness: $\|F(u)-F(v)\| \le L\|u-v\|$
- Unbiasedness: $\mathbb{E}[\tilde{F}(z)] = F(z)$
- Variance bound: $\mathbb{E}\|\tilde{F}(z)-F(z)\|^2 \le R^2$

Theorem (Deterministic). [Korpelevich, 1976, Nemirovski, 2004]

EG w/ $\tilde{F} = F$, $\boxed{\eta_t = \eta < 1/L}$:

- Convergence: $z_{t+1} \to z^\star$
- Rate[2]: $\mathrm{err}\left(\frac{1}{T}\sum_{t=1}^T z_{t+1/2}\right) \le \mathcal{O}\left(\frac{1}{T}\right)$

Theorem (Stochastic). [Juditsky et al., 2011]

EG w/ noisy oracles $\mathbb{E}[\tilde{F}(z)] = F(z)$, $\boxed{\eta_t = \frac{\eta_0}{\sqrt{t}}}$:

- Rate: $\mathbb{E}\,\mathrm{err}\left(\frac{1}{T}\sum_{t=1}^T z_{t+1/2}\right) \le \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$

[1] Same ideas also apply when we use Bregman distances in the update rule. This version is known as Mirror-Prox.

[2] Merit function for the rate is primal-dual gap: $\mathrm{err}(\bar{z}) = \max_{x,y}\Phi(\bar{x},y) - \Phi(x,\bar{y})$.

Takeaway

Algorithm (Stochastic) Extragradient

for $t = 0$ to $T - 1$ **do**
  $z_{t+1/2} = P_{\mathcal{Z}}(z_t - \eta_t \tilde{F}(z_t))$
  $z_{t+1} = P_{\mathcal{Z}}(z_t - \eta_t \tilde{F}(z_{t+1/2}))$
end for

$\rightarrow$ Different step sizes for deterministic & stochastic: $1/L$ vs $\eta_0/\sqrt{t}$

$\rightarrow$ Need to know L to set step size

Outline

→ Classic methods

Adaptivity to	Lipschitz constant	noise	strong convexity	sparsity of data
[Bach and Levy, 2019]	✓	✓	×	×
[Chambolle et al., 2018]	×	×	✓	×
[Alacaoglu et al., 2020]	×	×	✓	✓

Adaptivity to smoothness and noise

- $F(z) = (\nabla_x \Phi(x,y), -\nabla_y \Phi(x,y)) = \mathbb{E}[\tilde{F}(z)]$

Algorithm (Stochastic) Adaptive extragradient

---

for $t = 0$ to $T-1$ **do**

$z_{t+1/2} = P_{\mathcal{Z}}(z_t - \eta_t \tilde{F}(z_t))$

$z_{t+1} = P_{\mathcal{Z}}(z_t - \eta_t \tilde{F}(z_{t+1/2}))$

end for

- Run EG w/ adaptive step size [Bach and Levy, 2019] $(\max_{x,y} \|x-y\| \le D, G_0 > 0)$,

$$\eta_t = \frac{D}{\sqrt{G_0^2 + \sum_{i=0}^{t-1} Z_i^2}}, \tag{1}$$

where, $Z_i^2 = \frac{\|z_{i+1} - z_{i+1/2}\|^2 + \|z_{i+1/2} - z_i\|^2}{5\eta_i^2}$.

Adaptivity to smoothness and noise

- $F(z) = (\nabla_x \Phi(x,y), -\nabla_y \Phi(x,y)) = \mathbb{E}[\tilde{F}(z)]$

Algorithm (Stochastic) Adaptive extragradient

for $t = 0$ to $T-1$ **do**
  $z_{t+1/2} = P_{\mathcal{Z}}(z_t - \eta_t \tilde{F}(z_t))$
  $z_{t+1} = P_{\mathcal{Z}}(z_t - \eta_t \tilde{F}(z_{t+1/2}))$
end for

- Run EG w/ adaptive step size [Bach and Levy, 2019] ($\max_{x,y} \|x-y\| \leq D, G_0 > 0$),

$$\eta_t = \frac{D}{\sqrt{G_0^2 + \sum_{i=0}^{t-1} Z_i^2}}, \tag{1}$$

where, $Z_i^2 = \frac{\|z_{i+1} - z_{i+1/2}\|^2 + \|z_{i+1/2} - z_i\|^2}{5\eta_i^2}$.

Intuition. Recall AdaGrad step size for $\min_x f(x)$: $\eta_t = \frac{D}{\sqrt{\sum_{i=1}^{t} \|\nabla f(x_i)\|^2}}$.

- $Z_i^2 \sim \|\tilde{F}(z)\|^2$, since when $\mathcal{Z} = \mathbb{R}^{d+n}$,

$$\|z_{i+1/2} - z_i\| = \|\eta_i \tilde{F}(z_i)\|$$

Adaptivity to smoothness and noise

- $F(z) = (\nabla_x \Phi(x,y), -\nabla_y \Phi(x,y)) = \mathbb{E}[\tilde{F}(z)]$

Algorithm (Stochastic) Adaptive extragradient

---

for $t = 0$ to $T-1$ **do**
  $z_{t+1/2} = P_{\mathcal{Z}}(z_t - \eta_t \tilde{F}(z_t))$
  $z_{t+1} = P_{\mathcal{Z}}(z_t - \eta_t \tilde{F}(z_{t+1/2}))$
end for

- Run EG w/ adaptive step size [Bach and Levy, 2019] $(\max_{x,y} \|x-y\| \le D, G_0 > 0)$,

$$\eta_t = \frac{D}{\sqrt{G_0^2 + \sum_{i=0}^{t-1} Z_i^2}}, \tag{1}$$

where, $Z_i^2 = \frac{\|z_{i+1} - z_{i+1/2}\|^2 + \|z_{i+1/2} - z_i\|^2}{5\eta_i^2}$.

Theorem (Deterministic). [Bach and Levy, 2019]
EG w/ perfect oracle $\tilde{F} = F$ & η_t in (1):
- Rate: $\mathrm{err}\left(\frac{1}{T}\sum_{t=1}^{T} z_{t+1/2}\right) \le \mathcal{O}\left(\frac{1}{T}\right)$

Theorem (Stochastic). [Bach and Levy, 2019]
EG w/ noisy oracles $\mathbb{E}[\tilde{F}(z)] = F(z)$ & η_t in (1):
- Rate: $\mathbb{E}\,\mathrm{err}\left(\frac{1}{T}\sum_{t=1}^{T} z_{t+1/2}\right) \le \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$

Takeaway

Algorithm (Stochastic) Adaptive extragradient [Bach and Levy, 2019]

for $t = 0$ to $T - 1$ **do**
  $z_{t+1/2} = P_{\mathcal{Z}}(z_t - \eta_t \tilde{F}(z_t))$
  $z_{t+1} = P_{\mathcal{Z}}(z_t - \eta_t \tilde{F}(z_{t+1/2}))$
end for

- EG+AdaGrad step size

$$\eta_t = \frac{D}{\sqrt{G_0^2 + \sum_{i=0}^{t-1} Z_i^2}},$$

where, $Z_i^2 = \frac{\|z_{i+1} - z_{i+1/2}\|^2 + \|z_{i+1/2} - z_i\|^2}{5\eta_i^2}$.

→ Same step size for deterministic & stochastic

→ No need to know L

→ Optimal rate interpolation between for deterministic & stochastic

ExtraAdam

Algorithm ExtraAdam [Gidel et al., 2018]

for $t = 0$ to $T - 1$ **do**

$$\left.\begin{aligned} g_t &= \tilde{F}(z_t) \\ m_{t-1/2} &= \beta_1 m_{t-1} + (1-\beta_1) g_t \\ v_{t-1/2} &= \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \\ z_{t+1/2} &= z_t - \frac{\eta_t}{\sqrt{v_{t-1/2}}} m_{t-1} \end{aligned}\right\} \text{Extragradient step}^1$$

$$\left.\begin{aligned} g_{t+1/2} &= \tilde{F}(z_{t+1/2}) \\ m_t &= \beta_1 m_{t-1/2} + (1-\beta_1) g_{t+1/2} \\ v_t &= \beta_2 v_{t-1/2} + (1-\beta_2) g_{t+1/2}^2 \\ z_{t+1} &= z_t - \frac{\eta_t}{\sqrt{v_t}} m_t \end{aligned}\right\} \text{Main update step}$$

end for

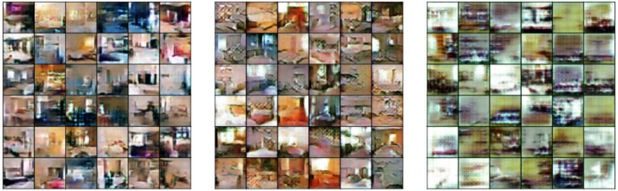
→ Extragradient + Adam

→ Limited theoretical understanding

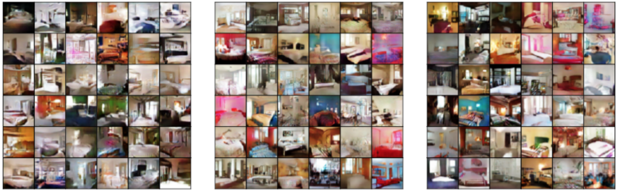
→ Compelling practical performance for training GANs

[1] Bias correction steps are omitted from ExtraAdam for simplicity.

Real LSUN Dataset: Extra-Adam, $4 \times 10^4, 8 \times 10^4, \times 10^5$ iterations [Hsieh et al., 2019]



(d) Simultaneous Extra-Adam



(e) Alternated Extra-Adam

Outline

→ Classic methods

Adaptivity to	Lipschitz constant	noise	strong convexity	sparsity of data
[Bach and Levy, 2019]	✓	✓	×	×
[Chambolle et al., 2018]	×	×	✓	×
[Alacaoglu et al., 2020]	×	×	✓	✓

Bilinear setting

$$\boxed{\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^n}\langle Ax, y\rangle + f(x) - h(y)}$$

- f, h are closed, convex functions
- Solution set is nonempty
- Linearly constrained problems, TV regularization, empirical risk minimization...

Algorithm Alternating GDA (Arrow-Hurwicz)

for $t = 0$ to $T-1$ **do**
  $x_{t+1} = \mathrm{prox}_{\tau f}(x_t - \tau A^\top y_t)$
  $y_{t+1} = \mathrm{prox}_{\sigma h}(y_t + \sigma A x_{t+1})$
end for

- $\mathrm{prox}_{\tau g}(u) = \arg\min_x g(x) + \frac{1}{2\tau}\|x-u\|^2$

Bilinear setting

$$\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^n} \langle Ax, y \rangle + f(x) - h(y)$$

- f, h are closed, convex functions
- Solution set is nonempty
- Linearly constrained problems, TV regularization, empirical risk minimization...

Algorithm Alternating GDA (Arrow-Hurwicz)

for $t = 0$ to $T-1$ **do**
  $x_{t+1} = \mathrm{prox}_{\tau f}(x_t - \tau A^\top y_t)$
  $y_{t+1} = \mathrm{prox}_{\sigma h}(y_t + \sigma A x_{t+1})$
end for

- $\mathrm{prox}_{\tau g}(u) = \arg\min_x g(x) + \frac{1}{2\tau}\|x-u\|^2$

Algorithm Primal-dual hybrid gradient (PDHG) [Chambolle and Pock, 2011]

for $t = 0$ to $T-1$ **do**
  $x_{t+1} = \mathrm{prox}_{\tau f}(x_t - \tau A^\top (y_t + y_t - y_{t-1}))$
  $y_{t+1} = \mathrm{prox}_{\sigma h}(y_t + \sigma A x_{t+1})$
end for

- PDHG w/ step sizes $\tau\sigma\|A\|^2 < 1$ converges to $(x^\star, y^\star)$ [Chambolle and Pock, 2011].
- Rate: $\mathrm{err}\left(\frac{1}{T}\sum_{t=1}^{T} z_t\right) \leq \mathcal{O}\left(\frac{1}{T}\right)$

Separable bilinear setting

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^n} f(x) + \sum_{i=1}^n \langle A_i x, y_i\rangle - h_i(y_i)$$

- Linearly constrained problems, TV regularization, empirical risk minimization...

Algorithm Stochastic Primal-dual hybrid gradient (SPDHG) [Chambolle et al., 2018]

for $t = 0$ to $T-1$ **do**
  $x_{t+1} = \mathrm{prox}_{\tau f}(x_t - \tau A^\top \bar{y}_t)$
  Pick $i_t \in [n]$ randomly
  $y_{t+1,i_t} = \mathrm{prox}_{\sigma_{i_t} h_{i_t}}(y_{t,i_t} + \sigma A_{i_t} x_{t+1})$
  $y_{t+1,i} = y_{t,i}$, for $i \neq i_t$
  $\bar{y}_{t+1} = y_{t+1} + n(y_{t+1} - y_t)$
end for

Separable bilinear setting

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^n} f(x) + \sum_{i=1}^{n}\langle A_i x, y_i\rangle - h_i(y_i)$$

- Linearly constrained problems, TV regularization, empirical risk minimization...

Algorithm Stochastic Primal-dual hybrid gradient (SPDHG) [Chambolle et al., 2018]

for $t = 0$ to $T-1$ **do**
  $x_{t+1} = \text{prox}_{\tau f}(x_t - \tau A^\top \bar{y}_t)$
  Pick $i_t \in [n]$ randomly
  $y_{t+1,i_t} = \text{prox}_{\sigma_{i_t} h_{i_t}}(y_{t,i_t} + \sigma A_{i_t} x_{t+1})$
  $y_{t+1,i} = y_{t,i}$, for $i \neq i_t$
  $\bar{y}_{t+1} = y_{t+1} + n(y_{t+1} - y_t)$
end for

Theorem. [Alacaoglu et al., 2019] PDHG w/ step sizes $n\tau\sigma_i\|A_i\|^2 < 1$:

- Convergence: $(x_t, y_t) \to (x^\star, y^\star)$
- Rate:
$\mathbb{E}\,\text{err}\left(\frac{1}{T}\sum_{t=1}^{T} x_t, \frac{1}{T}\sum_{t=1}^{T} y_t\right) \leq \mathcal{O}\left(\frac{1}{T}\right)$

Separable bilinear setting

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^n} f(x) + \sum_{i=1}^n \langle A_i x, y_i\rangle - h_i(y_i)$$

◦ Linearly constrained problems, TV regularization, empirical risk minimization...

Algorithm Stochastic Primal-dual hybrid gradient (SPDHG) [Chambolle et al., 2018]

for $t = 0$ to $T-1$ **do**
  $x_{t+1} = \text{prox}_{\tau f}(x_t - \tau A^\top \bar{y}_t)$
  Pick $i_t \in [n]$ randomly
  $y_{t+1,i_t} = \text{prox}_{\sigma_{i_t} h_{i_t}}(y_{t,i_t} + \sigma A_{i_t} x_{t+1})$
  $y_{t+1,i} = y_{t,i}$, for $i \neq i_t$
  $\bar{y}_{t+1} = y_{t+1} + n(y_{t+1} - y_t)$
end for

Theorem. [Alacaoglu et al., 2019] PDHG w/ step sizes $n\tau\sigma_i\|A_i\|^2 < 1$:

◦ Convergence: $(x_t, y_t) \to (x^\star, y^\star)$

◦ Rate:
$\mathbb{E}\,\text{err}\left(\frac{1}{T}\sum_{t=1}^T x_t, \frac{1}{T}\sum_{t=1}^T y_t\right) \le \mathcal{O}\left(\frac{1}{T}\right)$

Theorem. [Chambolle et al., 2018] If f and h are (μ_f, μ_i) strongly convex and τ, σ_i are chosen depending on (μ_f, μ_i):

(x_t, y_t) converge linearly to $(x^\star, y^\star)$.

Takeaway

$$\min_{x\in\mathbb{R}^d} \max_{y\in\mathbb{R}^n} f(x) + \sum_{i=1}^{n} \langle A_i x, y_i \rangle - h_i(y_i)$$

Algorithm Stochastic Primal-dual hybrid gradient (SPDHG)

for $t = 0$ to $T-1$ **do**
  $x_{t+1} = \text{prox}_{\tau f}(x_t - \tau A^\top \bar{y}_t)$
  Pick $i_t \in [n]$ randomly
  $y_{t+1,i_t} = \text{prox}_{\sigma_{i_t} h_{i_t}}(y_{t,i_t} + \sigma A_{i_t} x_{t+1})$
  $y_{t+1,i} = y_{t,i}$, for $i \neq i_t$
  $\bar{y}_{t+1} = y_{t+1} + n(y_{t+1} - y_t)$
end for

→ Different step sizes for sublinear & linear rate

→ Need knowledge of μ_f, μ_i to set step sizes for linear rate

Convergence of SPDHG under metric subregularity

$$\min_{x\in\mathbb{R}^d} \max_{y\in\mathbb{R}^n} f(x) + \sum_{i=1}^n \langle A_i x, y_i\rangle - h_i(y_i)$$

Algorithm Stochastic Primal-dual hybrid gradient (SPDHG)

for $t = 0$ to $T-1$ **do**
 $x_{t+1} = \text{prox}_{\tau f}(x_t - \tau A^\top \bar{y}_t)$
 Pick $i_t \in [n]$ randomly
 $y_{t+1,i_t} = \text{prox}_{\sigma_{i_t} h_{i_t}}(y_{t,i_t} + \sigma A_{i_t} x_{t+1})$
 $y_{t+1,i} = y_{t,i}$, for $i \neq i_t$
 $\bar{y}_{t+1} = y_{t+1} + n(y_{t+1} - y_t)$
end for

◦ Metric subregularity: Generalization of strong convexity.

◦ Satisfied when f, h are strongly convex, or when f, h are piecewise linear-quadratic (PLQ): indicator of polyhedral sets, polyhedral norms, hinge loss, Huber loss etc.

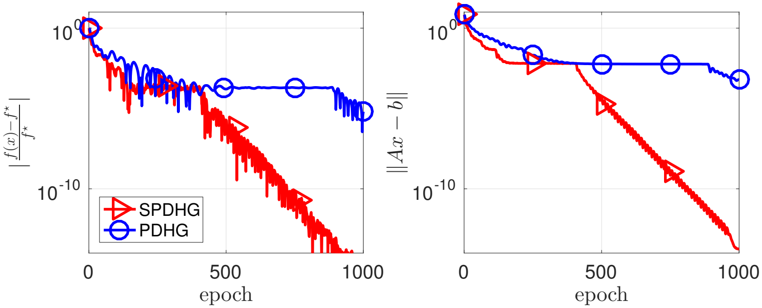
Theorem. [Alacaoglu et al., 2019] Assume metric subregularity and pick τ, σ as $n\tau\sigma_i\|A_i\|^2 < 1$, then

(x_t, y_t) converge linearly to $(x^\star, y^\star)$.

Performance

$$\min_{x \in \mathbb{R}^d} \|x\|_1 : \quad Ax = b$$

- Synthetic setup: A has normal distribution with $\Sigma_{i,j} = 0.5^{|i-j|}$
- $n, d = 500, 1000$



Takeaway

Algorithm Stochastic Primal-dual hybrid gradient (SPDHG) [Chambolle et al., 2018]

for $t = 0$ to $T - 1$ **do**
  $x_{t+1} = \text{prox}_{\tau f}(x_t - \tau A^\top \bar{y}_t)$
  Pick $i_t \in [n]$ randomly
  $y_{t+1,i_t} = \text{prox}_{\sigma_{i_t} h_{i_t}}(y_{t,i_t} + \sigma A_{i_t} x_{t+1})$
  $y_{t+1,i} = y_{t,i}$, for $i \neq i_t$
  $\bar{y}_{t+1} = y_{t+1} + n(y_{t+1} - y_t)$
end for

→ Randomization speeds up PDHG

→ Same step sizes with and without strong convexity

→ No need to know μ_f, μ_i to obtain linear convergence

◦ Our main reference for theoretical results: [Alacaoglu et al., 2019]

Outline

→ Classic methods

Adaptivity to	Lipschitz constant	noise	strong convexity	sparsity of data
[Bach and Levy, 2019]	✓	✓	×	×
[Chambolle et al., 2018]	×	×	✓	×
[Alacaoglu et al., 2020]	×	×	✓	✓

Per iteration cost of SPDHG

$$\min_{x\in\mathbb{R}^d} \max_{y\in\mathbb{R}^n} f(x) + \sum_{i=1}^{n} \langle A_i x, y_i \rangle - h_i(y_i)$$

Algorithm Stochastic Primal-dual hybrid gradient (SPDHG)

for $t = 0$ to $T-1$ **do**

$x_{t+1} = \mathrm{prox}_{\tau f}(x_t - \tau A^\top \bar{y}_t)$

Pick $i_t \in [n]$ randomly

$y_{t+1,i_t} = \mathrm{prox}_{\sigma_{i_t} h_{i_t}}(y_{t,i_t} + \sigma A_{i_t} x_{t+1})$

$y_{t+1,i} = y_{t,i}$, for $i \neq i_t$

$\bar{y}_{t+1} = y_{t+1} + n(y_{t+1} - y_t)$

end for

Per iteration cost of SPDHG

$$\min_{x\in\mathbb{R}^d} \max_{y\in\mathbb{R}^n} f(x) + \sum_{i=1}^{n} \langle A_i x, y_i \rangle - h_i(y_i)$$

Algorithm Stochastic Primal-dual hybrid gradient (SPDHG)

for $t = 0$ to $T-1$ **do**
- $x_{t+1} = \mathrm{prox}_{\tau f}(x_t - \tau A^\top \bar{y}_t)$
- Pick $i_t \in [n]$ randomly
- $y_{t+1,i_t} = \mathrm{prox}_{\sigma_{i_t} h_{i_t}}(y_{t,i_t} + \sigma A_{i_t} x_{t+1})$
- $y_{t+1,i} = y_{t,i}$, for $i \neq i_t$
- $\bar{y}_{t+1} = y_{t+1} + n(y_{t+1} - y_t)$

end for

Analysis of per iteration cost:

- Recall $A = \begin{bmatrix} A_1 \\ \vdots \\ A_n \end{bmatrix}$, with $A_i^\top \in \mathbb{R}^d$
- Compute $x_{t+1} \to$ cost d.
- $[Ax_{k+1}]_{i_t} = \langle A_{i_t}^\top, x_{k+1} \rangle \to$ cost $\mathrm{nnz}(A_{i_t})$.
- We maintain $A^\top y_t$ and compute:

$$A^\top \bar{y}_{t+1} = A^\top y_t + (n+1) A_{i_t}^\top (y_{t+1}^{i_t} - y_t^{i_t})$$

$\to$ cost $\mathrm{nnz}(A_{i_t})$.

PURE-CD

Algorithm PURE-CD [Alacaoglu et al., 2020]

for $t = 0$ to $T-1$ **do**

$\bar{x}_{t+1} = \text{prox}_{\tau f}(x_t - \tau A^\top y_t)$

$\bar{y}_{t+1} = \text{prox}_{\sigma h}(y_t + \sigma A \bar{x}_{t+1})$

Pick $i_t \in [n]$ randomly

$y_{t+1,i_t} = \bar{y}_{t+1,i_t}$

$y_{t+1,j} = y_{t,j}, \forall j \neq i_t$

$x_{t+1,j} = \bar{x}_{t+1,j} - \tau_j \theta_j [A^\top (y_{t+1} - y_t)]_j, \forall j \in J(i_t)$

$x_{t+1,j} = x_{t,j}, \forall j \notin J(i_t)$

end for

Parameters:

- $I(j) = \{i \in [n] : A_{i,j} \neq 0\}$
- $J(i) = \{j \in [d] : A_{i,j} \neq 0\}$
- $(\tau_i)_{i=1}^d$ and $(\sigma_i)_{i=1}^n$ are chosen using nonzero entries of A due to $I(j)$ and $J(i)$.

step size w. dense A	iter. cost
$n\tau\sigma_i \|A_i\|^2 < 1$	$\text{nnz}(A_i)$

... compared to SPDHG that had

step size w. dense A	iter. cost
$n\tau\sigma_i \|A_i\|^2 < 1$	d

→ Significant improvement when d is big & A is sparse

→ Similar theoretical rates as SPDHG

PURE-CD

- Lasso with different levels of sparsity of data.

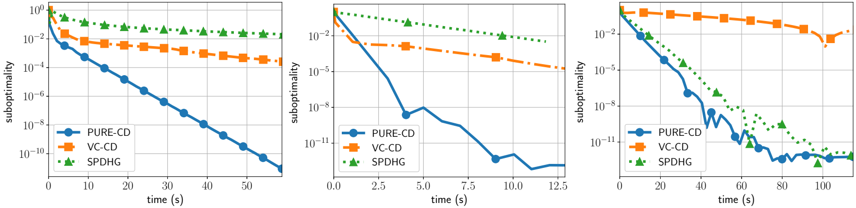


Figure: Lasso: Left: rcv1, $n = 20,242, m = 47,236$, density $= 0.16\%$, $\lambda = 10$; Middle: w8a, $n = 49,749, m = 300$, density $= 3.9\%$, $\lambda = 10^{-1}$; Right: covtype, $n = 581,012$, $m = 54$, density $= 22.1\%$, $\lambda = 10$.

Takeaway

Algorithm PURE-CD [Alacaoglu et al., 2020]

for $t = 0$ to $T-1$ **do**

$\bar{x}_{t+1} = \text{prox}_{\tau f}(x_t - \tau A^\top y_t)$

$\bar{y}_{t+1} = \text{prox}_{\sigma h}(y_t + \sigma A \bar{x}_{t+1})$

Pick $i_t \in [n]$ randomly

$y_{t+1,i_t} = \bar{y}_{t+1,i_t}$

$y_{t+1,j} = y_{t,j}, \forall j \neq i_t$

$x_{t+1,j} = \bar{x}_{t+1,j} - \tau_j \theta_j [A^\top (y_{t+1} - y_t)]_j, \forall j \in J(i_t)$

$x_{t+1,j} = x_{t,j}, \forall j \notin J(i_t)$

end for

→ Randomization speeds up PDHG

→ No need to know μ_f, μ_i to obtain linear convergence

→ Per-iteration cost and step sizes adapt to sparsity.

Summary

$$\boxed{\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \Phi(x,y)}$$

Adaptivity to	Lipschitz constant	noise	strong convexity	sparsity of data
[Bach and Levy, 2019]	✓	✓	×	×
[Chambolle et al., 2018]	×	×	✓	×
[Alacaoglu et al., 2020]	×	×	✓	✓

References I

Alacaoglu, A., Fercoq, O., and Cevher, V. (2019).
On the convergence of stochastic primal-dual hybrid gradient.
arXiv preprint arXiv:1911.00799.

Alacaoglu, A., Fercoq, O., and Cevher, V. (2020).
Random extrapolation for primal-dual coordinate descent.
In *International conference on machine learning*, pages 191–201. PMLR.

Arrow, K. J., Azawa, H., Hurwicz, L., and Uzawa, H. (1958).
Studies in linear and non-linear programming, volume 2.
Stanford University Press.

Bach, F. and Levy, K. Y. (2019).
A universal algorithm for variational inequalities adaptive to smoothness and noise.
In *Conference on Learning Theory*, pages 164–194.

Chambolle, A., Ehrhardt, M. J., Richtárik, P., and Schonlieb, C.-B. (2018).
Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications.
SIAM Journal on Optimization, 28(4):2783–2808.

References II

Chambolle, A. and Pock, T. (2011).
A first-order primal-dual algorithm for convex problems with applications to imaging.
Journal of Mathematical Imaging and Vision, 40(1):120–145.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. (2018).
A variational inequality perspective on generative adversarial networks.
In *International Conference on Learning Representations*.

Hsieh, Y.-P., Liu, C., and Cevher, V. (2019).
Finding mixed Nash equilibria of generative adversarial networks.
volume 97 of *Proceedings of Machine Learning Research*, pages 2810–2819, Long Beach, California, USA. PMLR.

Juditsky, A., Nemirovski, A., and Tauvel, C. (2011).
Solving variational inequalities with stochastic mirror-prox algorithm.
Stochastic Systems, 1(1):17–58.

Korpelevich, G. M. (1976).
The extragradient method for finding saddle points and other problems.
Matecon, 12:747–756.

References III

Nemirovski, A. (2004).
Prox-method with rate of convergence O(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems.
SIAM Journal on Optimization, 15(1):229–251.

Sibony, M. (1970).
Méthodes itératives pour les équations et inéquations aux dérivées partielles non linéaires de type monotone.
Calcolo, 7(1-2):65–183.