CS794/CO673: Optimization for Data Science

Lec 20: Randomized Smoothing

Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

November 25, 2022

Problem

Constrained minimization:

$$\min_{\mathbf{w} \in C \subseteq \mathbb{R}^d} \quad f(\mathbf{w})$$

- C is closed convex and f is (non)convex
- Can only evaluate the function value $f(\mathbf{w})$ but not the (sub)gradient
- Zero-th order method (a.k.a. gradient-free or derivative-free)
- For most (if not all) functions in practice, computing the function value (a scalar) costs as much as computing a (sub)gradient (a vector)!
- But only when we have *direct* access to the inner workings of f

$$\min_{x \in [a,b]} \quad f(x)$$

Algorithm 1: Golden-section search

Input: $a < b, g = \frac{\sqrt{5}+1}{2}$, `tol`

```
x_1 = a + (b - a)/g
x_2 = b - (b - a)/g
while x_2 - x_1 > tol do
    if f(x_2) > f(x_1) then
        b = x_2
        x_2 = a + (b - a)/g
    else
        a = x_1
        x_1 = b - (b - a)/g
```

Fix the number of evaluations. Is there an "optimal" alg?

$$\inf_{\mathcal{A}} \sup_{f} \text{ length of returned interval}$$

Key idea: recycle!

$$\min_{\lambda_2 \leq 1/2} \prod_{i=2}^{N} (1 - \lambda_i), \qquad \text{s.t.} \quad \lambda_{n+1} = \frac{\lambda_n}{1 - \lambda_n} \wedge \frac{1 - 2\lambda_n}{1 - \lambda_n}$$

Solution: $\lambda_n = \frac{F_{n-1}}{F_{n+1}}$

J. Kiefer. "Sequential Minimax Search for a Maximum". *Proceedings of the American Mathematical Society*, vol. 4, no. 3 (1953), pp. 502–506.

Uniform Grid Search

Algorithm 2: Random pursuit

Input: $\mathbf{w}_0$ such that $[\![f \leq f(\mathbf{w}_0)]\!]$ is compact

```
for t = 1, 2, ... do
    choose normalized direction d_t randomly
    η_t ← argmin_{η∈ℝ} f(w_t + η d_t)        // line search on chosen direction
    w_{t+1} ← w_t + η_t d_t
```

S. U. Stich et al. "Optimization of Convex Functions with Random Pursuit". *SIAM Journal on Optimization*, vol. 23, no. 2 (2013), pp. 1284–1309, S. U. Stich et al. "Variable metric random pursuit". *Mathematical Programming*, vol. 156 (2016), pp. 549–579.

If f is L_1-smooth, then

$$\begin{aligned} f(\mathbf{w}_t + \eta_t \mathbf{d}_t) &\le f(\mathbf{w}_t) + \eta_t \langle \mathbf{d}_t, \nabla f(\mathbf{w}_t) \rangle + \tfrac{\mathsf{L}_1}{2}\eta_t^2 \\ &\le f(\mathbf{w}_t) + \eta(\mathbf{w} - \mathbf{w}_t)^\top \mathbf{d}_t \mathbf{d}_t^\top \nabla f(\mathbf{w}_t) + \tfrac{\mathsf{L}_1}{2}\eta^2 (\mathbf{w} - \mathbf{w}_t)^\top \mathbf{d}_t \mathbf{d}_t^\top (\mathbf{w} - \mathbf{w}_t)^\top \end{aligned}$$

- The above inequality is due to setting $\eta_t = \eta(\mathbf{w} - \mathbf{w}_t)^\top \mathbf{d}_t$ for some $\eta > 0$
- Using $\mathbb{E}\mathbf{d}_t\mathbf{d}_t^\top = \frac{1}{d}\mathbb{I}$ and assuming f is convex:

$$\begin{aligned} \mathbb{E} f(\mathbf{w}_t + \eta_t \mathbf{d}_t) &\le f(\mathbf{w}_t) + \tfrac{\eta}{d} \langle \mathbf{w} - \mathbf{w}_t, \nabla f(\mathbf{w}_t) \rangle + \tfrac{\eta^2 \mathsf{L}_1}{2d} \|\mathbf{w} - \mathbf{w}_t\|_2^2 \\ &\le f(\mathbf{w}_t) + \tfrac{\eta}{d} [f(\mathbf{w}) - f(\mathbf{w}_t)] + (\tfrac{\eta}{d})^2 \tfrac{d\mathsf{L}_1}{2} \|\mathbf{w} - \mathbf{w}_t\|_2^2 \end{aligned}$$

- A simple induction (as in conditional gradient) yields:

$$\mathbb{E}[f(\mathbf{w}_t) - f(\mathbf{w})] \le O\left(\tfrac{d\mathsf{L}_1}{t+1}\right)$$

- A factor of dimension d worse

Convolution

Definition: Convolution and Fourier transform

The convolution of two functions f and g is defined through integration:

$$(f * g)(\mathbf{w}) := \int_{\mathbf{z}} f(\mathbf{w} - \mathbf{z}) g(\mathbf{z}) \, \mathrm{d}\mathbf{z} = \int_{\mathbf{z}} f(\mathbf{z}) g(\mathbf{w} - \mathbf{z}) \, \mathrm{d}\mathbf{z} =: (g * f)(\mathbf{w}).$$

Recall the Fourier transform and its inverse:

$$(\mathscr{F} f)(\mathbf{w}^*) = \mathscr{F} f(\mathbf{w}^*) = \int_{\mathbf{w}} \exp(-2\pi i \langle \mathbf{w}, \mathbf{w}^* \rangle) f(\mathbf{w}) \, \mathrm{d}\mathbf{w}$$

$$(\mathscr{F}^{-1} g)(\mathbf{w}) = \int_{\mathbf{w}^*} \exp(2\pi i \langle \mathbf{w}, \mathbf{w}^* \rangle) g(\mathbf{w}^*) \, \mathrm{d}\mathbf{w}^*$$

$$\mathscr{F}(f*g)=\mathscr{F}f\cdot\mathscr{F}g,\quad \mathscr{F}\mathscr{F}^{-1}=\mathscr{F}^{-1}\mathscr{F}=\mathrm{Id},\quad \mathscr{F}f^{(\mathbf{k})}=(-2\pi i\mathbf{w}^*)^{\mathbf{k}}\mathscr{F}f$$

- Applying Fourier transform to the derivative of convolution:

$$\begin{aligned}\mathscr{F}(f*g)^{(\mathbf{k})}=(-2\pi i\mathbf{w}^*)^{\mathbf{k}}\cdot\mathscr{F}(f*g)=[(-2\pi i\mathbf{w}^*)^{\mathbf{k}}\mathscr{F}f]\mathscr{F}g&=\mathscr{F}(f^{(\mathbf{k})}*g)\\&=\mathscr{F}(f*g^{(\mathbf{k})})\end{aligned}$$

- Applying the inverse transform we obtain the formula of differentiating under the integral:

$$(f*g)^{(\mathbf{k})}=f^{(\mathbf{k})}*\mathbf{g}=f*g^{(\mathbf{k})}$$

- This can in fact be the definition of the derivative (distribution) of f, using the derivative of some super smooth functions g!

Randomized Smoothing

Definition:

For a (vector-valued) function $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^c$ we define its randomized smoothing as

$$\mathbf{f}_\gamma(\mathbf{w}) = \mathbb{E}\mathbf{f}(\mathbf{w} + \gamma\boldsymbol{\varepsilon}) = \mathbb{E}\mathbf{f}(\mathbf{w} - \gamma\boldsymbol{\varepsilon}),$$

where $\boldsymbol{\varepsilon}$ is some symmetric random noise with zero mean and identity covariance.

- Let p be the probability density function (pdf) of $\boldsymbol{\varepsilon}$
- Dilated density: $p_\gamma(\mathbf{z}) = \frac{1}{\gamma^d} p(\frac{1}{\gamma}\mathbf{z})$
- We have point-wise convergence:

$$\mathbf{f}_\gamma = \mathbb{E}\mathbf{f}(\mathbf{w} - \gamma\boldsymbol{\varepsilon}) = \mathbf{f} * p_\gamma, \text{ hence } \mathbf{f}_\gamma \to \mathbf{f} \text{ as } \gamma \to 0$$

- Intuitively expected, as the noise shrinks to 0, i.e. $p_\gamma \to \delta'_{\mathbf{0}}$

Calculus for Randomized Smoothing

- The map $\mathbf{f} \mapsto \mathbf{f}_\gamma$ is linear
- If f is convex/concave, so is f_γ
- If f is convex, then $f_\gamma \geq f$
- If $\mathbf{f}$ is L_0-Lipschitz continuous (w.r.t. $\|\cdot\|_2$ say), so is $\mathbf{f}_\gamma$. Moreover,

$$\|\mathbf{f}_\gamma - \mathbf{f}\|_2 \leq \gamma \mathsf{L}_0 \mathbb{E}\|\boldsymbol{\varepsilon}\|_2 \leq \gamma \mathsf{L}_0 \sqrt{\mathbb{E}\|\boldsymbol{\varepsilon}\|_2^2} = \gamma \mathsf{L}_0 \sqrt{d}$$

- If f is L_1-smooth (w.r.t. $\|\cdot\|_2$ say), so is f_γ. Moreover,

$$f_\gamma - f \leq \frac{\gamma^2 \mathsf{L}_1}{2} \mathbb{E}\|\boldsymbol{\varepsilon}\|_2^2 = \frac{\gamma^2 \mathsf{L}_1 d}{2},$$

  whereas a two-sided bound holds if both $\pm f$ are L_1-smooth.

Gradient approximation

- If $\pm f$ is L_1-smooth, then $\|\nabla f_\gamma - \nabla f\|_\circ \leq \gamma \mathsf{L}_1 \sqrt{d}$.
  - in fact, $\nabla f_\gamma = (\nabla f)_\gamma$, and $\|\nabla f\|_\circ \leq \|\nabla f_\gamma\|_\circ + \gamma \mathsf{L}_1 \sqrt{d}$
- If $\pm f$ is L_2-smooth, then $\|\nabla f_\gamma - \nabla f\|_\circ \leq \gamma^2 \mathsf{L}_2 d/2$.
  - in fact, $\nabla f_\gamma = (\nabla f)_\gamma$ and $\nabla^2 f_\gamma = (\nabla^2 f)_\gamma$

Justifying the Name

Differentiating under the integral we obtain

$$f_\gamma^{(\mathbf{k})} := [f * p_\gamma]^{(\mathbf{k})} = f^{(\mathbf{k}-\mathbf{l})} * p_\gamma^{(\mathbf{l})}, \quad \nabla^k f_\gamma(\mathbf{w}) = \int \nabla^{k-1} f(\mathbf{w}-\mathbf{z}) \otimes \nabla p_\gamma(\mathbf{z})\, \mathrm{d}\mathbf{z}.$$

Therefore, if f is L_{k-1}-smooth, then f_γ is L_k-smooth, where

$$\mathsf{L}_k \le \mathsf{L}_{k-1} \int \|\nabla p_\gamma(\mathbf{z})\|_2 \,\mathrm{d}\mathbf{z} = \frac{\mathsf{L}_{k-1}}{\gamma} \int \|\nabla p(\mathbf{z})\|_2 \,\mathrm{d}\mathbf{z} = \frac{s\mathsf{L}_{k-1}}{\gamma}$$

- $s := \mathbb{E}\|\nabla \ln p(\boldsymbol{\varepsilon})\|_2, \quad \boldsymbol{\varepsilon} \sim p$
- f_γ is (at least) 1 degree more smoother than f

$$
\begin{aligned}
\nabla f_\gamma(\mathbf{w}) = \int f(\mathbf{w}-\mathbf{z})\nabla p_\gamma(\mathbf{z})\,\mathrm{d}\mathbf{z} &= \tfrac{1}{\gamma}\mathbb{E}[f(\mathbf{w}-\gamma\boldsymbol{\varepsilon})\nabla \ln p(\boldsymbol{\varepsilon})] \\
&= -\tfrac{1}{\gamma}\mathbb{E}[f(\mathbf{w}+\gamma\boldsymbol{\varepsilon})\nabla \ln p(\boldsymbol{\varepsilon})] \\
&= -\mathbb{E}\left[\frac{f(\mathbf{w}+\gamma\boldsymbol{\varepsilon})-f(\mathbf{w})}{\gamma}\nabla \ln p(\boldsymbol{\varepsilon})\right] \\
&= -\mathbb{E}\left[\frac{f(\mathbf{w}+\gamma\boldsymbol{\varepsilon})-f(\mathbf{w}-\gamma\boldsymbol{\varepsilon})}{2\gamma}\nabla \ln p(\boldsymbol{\varepsilon})\right]
\end{aligned}
$$

When f is e.g. convex or an envelope function, we have the limit:

$$
\nabla f_0(\mathbf{w}) := -\mathbb{E}[f'(\mathbf{w};\boldsymbol{\varepsilon})\nabla \ln p(\boldsymbol{\varepsilon})], \quad \text{where} \quad f'(\mathbf{w};\boldsymbol{\varepsilon}) := \lim_{\gamma\downarrow 0}[f(\mathbf{w}+\gamma\boldsymbol{\varepsilon})-f(\mathbf{w})]/\gamma
$$

$$
= -\mathbb{E}[\sigma_{\partial f(\mathbf{w})}(\boldsymbol{\varepsilon})\nabla \ln p(\boldsymbol{\varepsilon})]
$$

Needless to say, when f is actually differentiable, we have $\nabla f_0 = \nabla f$.

Gaussian Smoothing

$$\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbb{I}), \quad i.e. \ \ p(\varepsilon) = (2\pi)^{d/2} \exp(-\|\varepsilon\|_2^2/2)$$

- $-\nabla \ln p(\varepsilon) = \varepsilon$ and $s = \mathbb{E}\|\nabla \ln p(\varepsilon)\|_2 \leq \sqrt{d}$
- Conveniently, f_γ is in fact infinitely many times differentiable, e.g.

$$\nabla f_\gamma(\mathbf{w}) = \frac{1}{\gamma}\mathbb{E}[f(\mathbf{w} + \gamma\varepsilon)\varepsilon] = \mathbb{E}\left[\frac{f(\mathbf{w}+\gamma\varepsilon)-f(\mathbf{w})}{\gamma}\varepsilon\right] = \mathbb{E}\left[\frac{f(\mathbf{w}+\gamma\varepsilon)-f(\mathbf{w}-\gamma\varepsilon)}{2\gamma}\varepsilon\right]$$

- Requires f to be defined on entire $\mathbb{R}^d$

Y. Nesterov and V. Spokoiny. "Random Gradient-Free Minimization of Convex Functions". *Foundations of Computational Mathematics*, vol. 17 (2017), pp. 527–566.

Uniform Smoothing

$$\boldsymbol{\varepsilon} \sim \mathrm{Uniform}(K), \quad i.e. \quad p(\boldsymbol{\varepsilon}) = \begin{cases} 1/v_d, & \text{if } \boldsymbol{\varepsilon} \in K \\ 0, & \text{otherwise} \end{cases}$$

- v_d is the volume of the (symmetric, isotropic, i.e. $\mathbb{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top = \mathbb{I}$) compact set K
- Applying Stokes' theorem, $\nabla p(\boldsymbol{\varepsilon}) = \mathbf{1}_{\partial K} \cdot \mathsf{n}(\boldsymbol{\varepsilon})/v_d$, where $\mathsf{n}(\boldsymbol{\varepsilon})$ is the normal vector
- $s = u_{d-1}/v_d$ where u_{d-1} is the surface area of ∂K; choose $\boldsymbol{\delta} \sim \mathrm{Uniform}(\partial K)$:

$$\nabla f_\gamma(\mathbf{w}) = -\tfrac{s}{\gamma}\mathbb{E}[f(\mathbf{w} + \gamma\boldsymbol{\delta})\mathsf{n}(\boldsymbol{\delta})] = -s\mathbb{E}\left[\tfrac{f(\mathbf{w}+\gamma\boldsymbol{\delta})-f(\mathbf{w})}{\gamma}\mathsf{n}(\boldsymbol{\delta})\right]$$
$$= -s\mathbb{E}\left[\tfrac{f(\mathbf{w}+\gamma\boldsymbol{\delta})-f(\mathbf{w}-\gamma\boldsymbol{\delta})}{2\gamma}\mathsf{n}(\boldsymbol{\delta})\right]$$

- Requires f to be defined (and bounded) over $C + \gamma K$.
- Let $K = \mathsf{B}_2(\mathbf{0}, \sqrt{d})$ we have $\mathsf{n}(\boldsymbol{\delta}) = -\sqrt{d}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2$ and $s = \sqrt{d}$

A. S. Nemirovski and D. B. Yudin. "**Problem complexity and method efficiency in optimization**". Wiley, 1983, A. D. Flaxman et al. "Online convex optimization in the bandit setting: gradient descent without a gradient". In: *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*. 2005, pp. 385–394.

Put Everything Together

- We optimize f_γ as a smoothed approximation of f
- We compute an unbiased, stochastic (sub)gradient of f_γ by
  1. $\hat{\partial}^1 f_\gamma(\mathbf{w}) = -\frac{1}{\gamma} f(\mathbf{w} + \gamma\boldsymbol{\epsilon}) \cdot \nabla \ln p(\boldsymbol{\epsilon})$
  2. $\hat{\partial}^{1,0} f_\gamma(\mathbf{w}) = -\frac{f(\mathbf{w}+\gamma\boldsymbol{\varepsilon})-f(\mathbf{w})}{\gamma} \cdot \nabla \ln p(\boldsymbol{\varepsilon})$
  3. $\hat{\partial}^{1,1} f_\gamma(\mathbf{w}) = -\frac{f(\mathbf{w}+\gamma\boldsymbol{\varepsilon})-f(\mathbf{w}-\gamma\boldsymbol{\varepsilon})}{2\gamma} \cdot \nabla \ln p(\boldsymbol{\varepsilon})$
  4. $\hat{\partial} f_0(\mathbf{w}) = -f'(\mathbf{w}; \boldsymbol{\varepsilon}) \cdot \nabla \ln p(\boldsymbol{\varepsilon})$
- Eexcept the last choice, only require 1 or 2 evaluations of the function
- Except the last choice, these stochastic (sub)gradients in general are biased for f
- We bound the second moment of the stochastic (sub)gradient
- We apply the stochastic GDA algorithm and obtain convergence towards f_γ
- We set γ appropriately so that we obtain convergence towards f

L_0-Lipschitz Continuous and Convex

- If f is convex, then $f_\gamma \geq f$
- If f is L_0-Lipschitz continuous (w.r.t. $\|\cdot\|_2$ say), so is $\mathbf{f}_\gamma$. Moreover,

$$\|\mathbf{f}_\gamma - \mathbf{f}\|_2 \leq \gamma \mathsf{L}_0 \mathbb{E}\|\boldsymbol{\varepsilon}\|_2 \leq \gamma \mathsf{L}_0 \sqrt{\mathbb{E}\|\boldsymbol{\varepsilon}\|_2^2} = \gamma \mathsf{L}_0 \sqrt{d}$$

- Thus, we obtain the approximation bound:

$$\mathbb{E}[f(\bar{\mathbf{w}}_t) - f(\mathbf{w})] - \gamma \mathsf{L}_0 \sqrt{d} \leq \mathbb{E}[f_\gamma(\bar{\mathbf{w}}_t) - f_\gamma(\mathbf{w})]$$

- Using $\hat{\partial}^{1,0} f_\gamma$ we obtain

$$\mathbb{E}[f_\gamma(\bar{\mathbf{w}}_t) - f_\gamma(\mathbf{w})] \leq \frac{\|\mathbf{w}_0 - \mathbf{w}\|_2^2 + \sum_{k=0}^{t} \eta_k^2 \cdot \mathbb{E}\|\hat{\partial}^{1,0} f_\gamma(\mathbf{w})\|_2^2}{2H_t}$$

- If f is L_0-Lipschitz continuous, then using Gaussian smoothing:

$$\begin{aligned}\mathbb{E}\|\hat{\partial}^{1,0} f_\gamma(\mathbf{w})\|_2^2 &= \mathbb{E}\left\|-\frac{f(\mathbf{w}+\gamma\boldsymbol{\varepsilon})-f(\mathbf{w})}{\gamma}\cdot\nabla\ln p(\boldsymbol{\varepsilon})\right\|_2^2 \\ &\le \mathsf{L}_0^2\cdot\mathbb{E}\|\boldsymbol{\varepsilon}\|_2^4 \\ &\le \mathsf{L}_0^2\cdot d(d+2)\le \mathsf{L}_0^2(d+1)^2\end{aligned}$$

- Setting $\gamma = \frac{\epsilon}{2\mathsf{L}_0\sqrt{d}}$, $\eta_t = \frac{\mathrm{diam}(C)}{(d+1)\mathsf{L}_0\sqrt{t+1}}$ we have

$$\mathbb{E}[f(\bar{\mathbf{w}}_t) - f(\mathbf{w})] \le \epsilon, \quad \text{if } t > \frac{4(d+1)^2}{\epsilon^2}[\mathrm{diam}(C)\mathsf{L}_0]^2,$$

which is d^2 times slower than running subgradient directly on f.

L_1-smooth and convex

- If f is convex, then $f_\gamma \geq f$
- If f is L_1-smooth (w.r.t. $\|\cdot\|_2$ say), so is f_γ. Moreover,

$$f_\gamma - f \leq \frac{\gamma^2 \mathsf{L}_1}{2} \mathbb{E}\|\boldsymbol{\varepsilon}\|_2^2 = \frac{\gamma^2 \mathsf{L}_1 d}{2}$$

- Thus, we obtain the approximation bound:

$$\mathbb{E}[f(\bar{\mathbf{w}}_t) - f(\mathbf{w})] - \frac{\gamma^2 \mathsf{L}_1 d}{2} \leq \mathbb{E}[f_\gamma(\bar{\mathbf{w}}_t) - f_\gamma(\mathbf{w})]$$

- Using again $\hat{\partial}^{1,0} f_\gamma$ we obtain similarly

$$\mathbb{E}[f_\gamma(\bar{\mathbf{w}}_t) - f_\gamma(\mathbf{w})] \leq \frac{\|\mathbf{w}_0 - \mathbf{w}\|_2^2 + \sum_{k=0}^t \eta_k^2 \cdot \mathbb{E}\|\hat{\partial}^{1,0} f_\gamma(\mathbf{w}_k)\|_2^2}{2H_t}$$

- If ∇f is L_1-Lipschitz continuous:

$$\begin{aligned}\mathbb{E}\|\hat{\partial}^{1,0} f_\gamma(\mathbf{w})\|_2^2 &= \mathbb{E}\left\|-\tfrac{f(\mathbf{w}+\gamma\boldsymbol{\varepsilon})-f(\mathbf{w})}{\gamma}\cdot\nabla\ln p(\boldsymbol{\varepsilon})\right\|_2^2 \\ &\le \mathbb{E}\left[\langle\nabla f(\mathbf{w}),\boldsymbol{\varepsilon}\rangle + \tfrac{\mathsf{L}_1\gamma\|\boldsymbol{\varepsilon}\|_2^2}{2}\right]^2\|\boldsymbol{\varepsilon}\|_2^2 \\ &\le \tfrac{\gamma^2\mathsf{L}_1^2}{2}d(d+2)(d+4)+2(d+2)\|\nabla f(\mathbf{w})\|_2^2\end{aligned}$$

- With $\gamma = O\big(\frac{1}{d}\sqrt{\frac{\epsilon}{\mathsf{L}_1}}\big)$ and $\eta_t \equiv O\big(\frac{1}{d\mathsf{L}_1}\big)$, need $O\big(\frac{d}{\epsilon}\mathsf{L}_1\,\mathrm{diam}^2(C)\big)$ many steps to obtain an ϵ-minimizer of f

- d times slower than running (projected) gradient directly on f

More Moment Bounds for Gaussian Smoothing

- If f is differentiable:

$$\mathbb{E}\|\hat{\partial} f_0(\mathbf{w})\|_2^2 = \mathbb{E}\|\boldsymbol{\varepsilon}\|_2^4 \left\langle \frac{\boldsymbol{\varepsilon}}{\|\boldsymbol{\varepsilon}\|_2}, \nabla f(\mathbf{w}) \right\rangle^2$$

$$= \mathbb{E}\|\boldsymbol{\varepsilon}\|_2^4 \cdot \mathbb{E} \left\langle \frac{\boldsymbol{\varepsilon}}{\|\boldsymbol{\varepsilon}\|_2}, \nabla f(\mathbf{w}) \right\rangle^2 = (d+2)\|\nabla f(\mathbf{w})\|_2^2$$

- If $\pm f$ is $\mathsf{L}_1^{\pm}$-smooth:

$$\mathbb{E}\|\hat{\partial}^{1,1} f_\gamma(\mathbf{w})\|_2^2 \leq \frac{\gamma^2(\mathsf{L}_1^+ + \mathsf{L}_1^-)^2}{8} d(d+2)(d+4) + 2(d+2)\|\nabla f(\mathbf{w})\|_2^2$$

- If $\nabla^2 f$ is L_2-Lipschitz continuous

$$\mathbb{E}\|\hat{\partial}^{1,1} f_\gamma(\mathbf{w})\|_2^2 \leq \frac{\gamma^4 \mathsf{L}_2^2}{18} d(d+2)(d+4)(d+6) + 2(d+2)\|\nabla f(\mathbf{w})\|_2^2$$